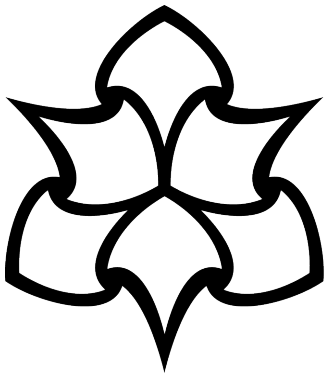


ABSTRACT

Background and Objective: Chronic pain following musculoskeletal trauma is common, which may partially be attributed to the early presence of central sensitisation (CS). Multiple measures are suggested to assess clinical features of CS, yet no systematic review has evaluated the measurement properties of these measures in a musculoskeletal trauma population.

Databases and Data Treatment: This systematic review, which followed a published and PROSPERO registered protocol (CRD42018091531), aimed to establish the scope of CS measures used within a musculoskeletal trauma population and evaluate their measurement properties. Searches were conducted in two stages by two independent reviewers. The Consensus-based Standards for the selection of Health Measurement instruments (COSMIN) checklist was used to evaluate risk of bias and overall quality was assessed using the modified Grading of Recommendations Assessment, Development and Evaluation.

Results: From 86 studies, 30 different CS outcome measures were identified. Nine studies evaluated measurement properties of nine outcome measures; eight evaluated reliability and one evaluated construct validity. Measures included seven quantitative sensory testing methods (pressure, cold and electrical pain thresholds; warm, cold and vibration detection thresholds; vibration perception thresholds), pain drawings and a pinwheel. Risk of bias was assessed as doubtful/inadequate for all but one study, overall quality of evidence was low/very low for all measures. Reliability of measures ranged from poor to excellent.

Conclusions: Many measures are used to evaluate CS but with limited established measurement properties in musculoskeletal trauma. High quality research to establish measurement properties of CS outcome measures is required.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/EJP.1670

Article type : Review Article

MEASURES OF CENTRAL SENSITISATION AND THEIR MEASUREMENT PROPERTIES IN MUSCULOSKELETAL TRAUMA: A SYSTEMATIC REVIEW

Middlebrook N[1,2], Rushton AB[1,2], Abichandani D[1,2], Kuithan P[1], Heneghan NR[1], Falla D[1,2]

Affiliations

[1] Centre of Precision Rehabilitation for Spinal Pain
School of Sport, Exercise and Rehabilitation Sciences
College of Life and Environmental Sciences
University of Birmingham
Edgbaston
Birmingham
B15 2TT
United Kingdom

[2] NIHR Surgical Reconstruction & Microbiology Research Centre
University of Birmingham
Edgbaston

Birmingham
B15 2TT
United Kingdom

Correspondence:

Professor Deborah Falla
Centre of Precision Rehabilitation for Spinal Pain
School of Sport, Exercise and Rehabilitation Sciences,
College of Life and Environmental Sciences,
University of Birmingham,
Birmingham, B15 2TT, UK.
Tel: +44 121 415 4220
Email: d.falla@bham.ac.uk

Category: Review

Funding: This study/project is funded by the National Institute for Health Research (NIHR) Surgical Reconstruction and Microbiology Research Centre (SRMRC). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Conflicts of Interest: The authors declare no conflict of interest

Significance: This systematic review is the first two staged review to collate measures evaluating central sensitisation within musculoskeletal trauma and evaluate the measurement properties of these measures within this population. This review highlights a mismatch between measures used and established measurement properties of central sensitisation measures within this population. This review is the first step towards a consensus on the most appropriate measures to use to evaluate central sensitisation in this population.

ABSTRACT

Background and Objective: Chronic pain following musculoskeletal trauma is common, which may partially be attributed to the early presence of central sensitisation (CS). Multiple measures are suggested to assess clinical features of CS, yet no systematic review has evaluated the measurement properties of these measures in a musculoskeletal trauma population.

Databases and Data Treatment: This systematic review, which followed a published and PROSPERO registered protocol (CRD42018091531), aimed to establish the scope of CS measures used within a musculoskeletal trauma population and evaluate their measurement properties. Searches were conducted in two stages by two independent reviewers. The Consensus-based Standards for the selection of Health Measurement instruments (COSMIN) checklist was used to evaluate risk of bias and overall quality was assessed using the modified Grading of Recommendations Assessment, Development and Evaluation.

Results: From 86 studies, 30 different CS outcome measures were identified. Nine studies evaluated measurement properties of nine outcome measures; eight evaluated reliability and one evaluated construct validity. Measures included seven quantitative sensory testing methods (pressure, cold and electrical pain thresholds; warm, cold and vibration detection thresholds; vibration perception thresholds), pain drawings and a pinwheel. Risk of bias was assessed as doubtful/inadequate for all but one study, overall quality of evidence was low/very low for all measures. Reliability of measures ranged from poor to excellent.

Conclusions: Many measures are used to evaluate CS but with limited established measurement properties in musculoskeletal trauma. High quality research to establish measurement properties of CS outcome measures is required.

INTRODUCTION

Chronic pain and poorer recovery are common following musculoskeletal trauma (Carroll et al., 2008; Rivara et al., 2008; Williamson et al., 2009). In chronic whiplash associated disorders (WAD), poorer outcome is linked to early widespread sensitisation (Sterling et al., 2003; Van Oosterwijck et al., 2013; Walton et al., 2011b), and following fractures, widespread pain distribution is evident soon after injury (Doménech-García et al., 2018). Due to noxious stimuli and prolonged duration of high pain intensity (Graven-Nielsen

and Arendt-Nielsen 2010; Katz and Seltzer 2009), hypersensitivity of the central nervous system or central sensitisation (CS) could occur in this population.

CS is defined as "increased responsiveness of nociceptive neurons in the central nervous system to their normal or subthreshold afferent input" (IASP 2017). CS can be generated and maintained by peripheral or central mechanisms, or a combination of both (Harte et al., 2018). This differs from the original definition of 'activity dependant CS' where changes occurred within the dorsal horn following a noxious peripheral stimulus (Woolf 2018). Multiple mechanisms including altered descending pain modulation, altered sensory processing within the brain, increase in glial activity and synaptic plasticity in spinal cord and cortex are suggested (Latremoliere and Woolf 2009; Nijs et al., 2019). Assessment of CS is challenging and clinical features of CS are often the focus due to the multiple mechanisms at play. Clinical features can include widespread pain, allodynia and secondary hyperalgesia (Latremoliere and Woolf 2009; Woolf 2011). However, assessing clinical features has its limitations in that this only can be suggestive of CS rather than a true diagnosis.

With complexity around CS, no gold standard measure exists (Neblett 2018). Multiple suggested methods to assess the features of CS include patient reported outcome measures (PROMS) such as the Central Sensitisation Inventory (Neblett 2018), Quantitative Sensory Testing (QST) (Cruz-Almeida and Fillingim 2014) and pain drawings to assess for widespread pain (Williams 2018).

Established measurement properties are required for outcome measures to avoid bias and to allow confidence in the research findings (Mokkink et al., 2010a). The Consensus-based Standards for the selection of Health Measurement instruments (COSMIN) initiative developed a consensus-based taxonomy of measurement properties to help improve the selection of outcome measures within health research (Mokkink et al., 2010b). Guidelines and a new tool to assess risk of bias for systematic reviews has been developed by COMSIN (Prinsen et al., 2018).

Previous systematic reviews summarising features of CS and outcome measurements used in WAD (Van Oosterwijck et al., 2013), and peripheral joint pain (Alqarni et al., 2018). highlight a wide range of outcome measures currently being utilised, particularly QST. Systematic reviews evaluating reliability of specific outcome measures such as thermal testing, (Moloney et al., 2012) and conditioned pain modulation (Kennedy et al., 2016) have been conducted. However, no systematic review has evaluated outcome measures used to

measure CS and their measurement properties within musculoskeletal trauma more generally to include other patient subgroups such as fractures and more major injuries. A systematic review is needed to synthesise current measures used within musculoskeletal trauma to allow a standardised approach and evaluate whether these measures have established measurement properties within musculoskeletal trauma populations. Therefore, the aims of this systematic review are:

1. Identify what outcome measures are used within musculoskeletal trauma research to evaluate presence of CS
2. Investigate whether current CS outcome measures used in musculoskeletal trauma research have established measurement properties.

LITERATURE SEARCH METHODS

Protocol and Registration

This systematic review followed a pre-defined published protocol (Middlebrook et al., 2019) which was prospectively registered with PROSPERO (CRD42018091531), and is reported in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Moher et al., 2009).

Eligibility Criteria

Inclusion Criteria

Population

Adults (aged >16 years) who experienced any type of musculoskeletal trauma were eligible. This included any musculoskeletal structure involved in a traumatic injury (Clay et al., 2010) and was inclusive of subgroups of traumatic injuries, including WAD, fracture, traumatic injuries involving amputation and gunshot or stab wounds. This definition of musculoskeletal trauma was carefully considered and pre-defined within our published protocol (Middlebrook et al., 2019), in the effort to be inclusive of all types of trauma in keeping with current management pathways within the United Kingdom (NICE 2018). For studies with a mixed population e.g. musculoskeletal trauma and traumatic brain injuries, the sample must have included more than 90% musculoskeletal trauma; a threshold adopted in previous systematic reviews within major musculoskeletal trauma (Clay et al., 2010; Clay et al., 2012).

Studies

Any study design apart from case studies, literature and systematic reviews were included. For any conference abstracts identified, authors were contacted to confirm whether the study had been published. No restriction on length of studies, time points or setting of study was observed.

Outcome Measures

Any outcome measure evaluating CS was included. A pre-defined criteria for defining CS was defined in the review protocol (Middlebrook et al., 2019). In brief, articles were included if they made reference to sensitisation of the central nervous system or reference to symptoms of sensitisation such as secondary hyperalgesia. Outcome measures could include any patient reported outcome measures designed to assess CS, performance-based measures such as QST or any measure used to evaluate symptoms of CS such as a pain drawing to measure widespread pain. This criterion was based on current literature and previous systematic reviews in this area of CS research (Clark et al., 2017; Fingleton et al., 2015; Latremoliere and Woolf 2009; Nijs et al., 2014; Vardeh et al., 2016; Woolf 2011).

Measurement Properties

Any domain or measurement property included in the COSMIN Taxonomy was included (Mokkink et al., 2018a; Mokkink et al., 2010b). The COSMIN taxonomy encompasses three main domains - reliability, validity and responsiveness, with each domain containing one or more measurement properties (Mokkink et al., 2010b).

Exclusion Criteria

Studies which investigated populations including traumatic brain injury, burns or neurological injury were excluded since established PROMs already exist for these subgroups. Studies where the full text was not written in English were excluded.

Information Sources

Two searches were conducted in this review: Stage 1 (inception to 23rd November 2018) and Stage 2 (inception to 21st July 2019). The following databases were included for both stages of the review:

- MEDLINE, EMBASE, CINAHL, ZETOC, Web of Science, PubMed, Google Scholar

- Hand Searching of key journals (Musculoskeletal Science and Practice, PAIN, European Journal of Pain, The Journal of Pain, The Clinical Journal of Pain)
- Grey literature searches including British National Bibliography, Open Grey, ProQuest, and EThOS.
- Leading authors in the field were contacted to ensure most up to date articles were obtained.

Search Strategy

This review was conducted in two stages:

1. Initial search (Stage 1) to identify the current CS measures used in the musculoskeletal trauma population.
2. Secondary search (Stage 2) to identify studies evaluating measurement properties of the measures identified in stage 1.

An example of the search strategy for MEDLINE for both stages can be found in the published protocol (Middlebrook et al., 2019). This search strategy was adapted for each database to allow for database search term variation.

Study Selection

For both stages two independent reviewers (Stage 1 NM/PK, Stage 2 NM/DA) searched information sources and reviewed titles and abstracts to evaluate whether articles were include/exclude/unsure based on the pre-defined eligibility criteria. Full text was sought for any articles which could not be excluded based on information in the abstract. Full text screening was conducted in the same independent manner. Articles were included if both reviewers agreed on eligibility. A third reviewer (AR, methodological expert) was consulted in the event of disagreement throughout the stages of the review. Authors were contacted if full texts were unavailable or, if further clarification was required on the study population. In the situation of two failed attempts to contact authors, an article was excluded.

Data collection process

Two reviewers (NM/DA) extracted data for both stages of the review using a standardised form. Authors were contacted if further information was required. The same procedure was followed whereby authors were contacted twice.

Data Items

Extracted data items from included studies for both review aims included study characteristics (study design, sample size, country of study), participant characteristics (age, gender, type and mechanism of trauma, duration of symptoms) and outcome measures (CS outcome measures and other outcome measures used). For stage two, additional data items specific to measurement properties were extracted including: time points, measurement property, statistical analyses and results.

Risk of bias in individual studies

The COSMIN Risk of Bias Checklist for systematic reviews (Mokkink et al., 2018a) was used to assess risk of bias in individual studies. The new COSMIN risk of bias checklist was used in this review, which was originally designed for PROMs. However, it is recommended that the tool can be adapted for other measures (Prinsen et al., 2018), and a pilot was conducted prior to the review to test the tool's suitability. Two reviewers (NM/DA) independently assessed risk of bias. The third reviewer (AR) was available in an event of disagreement between reviewers.

Summary Measures

For reliability statistics including intraclass correlation coefficients (ICCs) and Kappa coefficients, results and confidence intervals (95% CIs) are presented where possible.

Synthesis of results

Due to the heterogeneity of the studies (study population, outcome measures, data analysis), meta-analysis was not possible and therefore a narrative synthesis was conducted. Narrative synthesis was completed in line with the COSMIN guidelines for systematic reviews (Mokkink et al., 2018b). The COSMIN Risk of Bias checklist was completed and results of the studies was rated against the pre-defined criteria by COSMIN for good measurement properties (Mokkink et al., 2018b; Prinsen et al., 2018). The studies were then synthesised per outcome measure and per measurement property, and then rated overall against the criteria for good measurement properties (Mokkink et al., 2018b). Overall quality of evidence was then assessed using the modified Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach (Prinsen et al., 2018). The modified GRADE approach has been specifically adapted by COSMIN and, uses four of the five

GRADE factors: risk of bias, inconsistency, imprecision and indirectness (Prinsen et al., 2018). Inclusion of publication bias is not appropriate for this type of systematic review due to a lack of study register for measurement properties (Prinsen et al., 2018).

RESULTS

Study Selection

Stage 1

Fig. 1 summarises the articles included at each stage of the review. A total of 3330 articles were screened at title and abstract, with 142 assessed at full text stage. A total of 104 articles met the eligibility criteria. Twenty articles were identified in which the same dataset(s) of other articles identified eligible was used, with authors being contacted for further clarification when required, and subsequently excluded if the same dataset was used. However, two articles (Coppieters et al., 2017; Ris et al., 2018) used different outcome measures for CS and were therefore included. A total of 86 studies were included. Fig. 1 summarises the reasons for exclusion at full text stage.

FIGURE 1 INSERTED HERE

Stage 2

Fig. 2 summarises the articles included at each stage of the review. A total of 12,114 articles were screened at title and abstract stage, with 125 assessed at full text stage. For 29 articles, the population was unclear on whether it included more than 90% musculoskeletal trauma. Therefore, authors were contacted for further information. Of these 29, seven authors did not respond (Biurrun Manresa et al., 2011; Margolis et al., 1988; Margolis et al., 1986; Myburgh et al., 2011; O'Neill et al., 2014; Reigo et al., 1998; Vuilleumier et al., 2015), and for one article, the authors were not contactable (Cummings and Routan 1987). Therefore, we were unable to confirm the population for these articles and subsequently excluded them. In the instance where a conference abstract only was identified (n=6), authors were contacted to confirm if a full text was available. Three authors were not contactable (Christiansen et al., 2017; Cummings and Routon 1985; Fricton and Schiffman 1986), and one did not respond (Starz et al., 1995) and were therefore excluded. Two full texts were not available, and authors were not contactable (Blasco and Bayes 1988) or did not respond (Lahoda et al.,

1977) and they were not included. A total of nine studies were included, with Fig. 2 summarising the reasons for exclusion for the remaining studies.

FIGURE 2 INSERTED HERE

Study Characteristics

Stage 1

A summary of the CS measures used in all of the included studies is reported in Table S1. From the 30 different measures reported, three subcategories were derived - QST, PROMS and other. QST was the most frequently used of the subcategories. The most frequently used measurement of CS overall was pressure pain thresholds (PPT) using a handheld algometer (n=62). Cold pain thresholds (CPT) was the second most common method (n=32). Populations studied included, WAD (n=74), fractures (n=6), soft tissue injuries (n=2), foot and ankle trauma (n=1) and traumatic amputation (n=1). Descriptive data for the included studies for stage 1 can be found in Table S2.

Stage 2

Tables 1 and 2 summarise the study characteristics and results of the included nine studies. Six studies investigated a WAD population (Bock et al., 2005; Käll et al., 2008; Prushansky et al., 2007; Rushton et al., 2014; Southerst et al., 2013; Tyros et al., 2016). The remaining three studies investigated fractured wrists (Saebo et al., 2019), complex regional pain syndrome, (Kemler et al., 2000) and a mixed cohort of neck and shoulder pain (Bertilson et al., 2003). For the studies which evaluated complex regional pain syndrome and a mixed cohort of neck and shoulder pain, authors were contacted and they confirmed that their cohort included more than 90% musculoskeletal trauma and therefore were included in the review. Eight studies investigated reliability (Bertilson et al., 2003; Bock et al., 2005; Käll et al., 2008; Kemler et al., 2000; Prushansky et al., 2007; Saebo et al., 2019; Southerst et al., 2013; Tyros et al., 2016), one investigated validity (Rushton et al., 2014) and no studies evaluated responsiveness. CS measures investigated were pressure pain thresholds, (Prushansky et al., 2007; Saebo et al., 2019) pinwheel to evaluate allodynia and sensitivity to pain, (Bertilson et al., 2003; Bock et al., 2005) vibration perception (Rushton et al., 2014) and disappearance thresholds, (Rushton et al., 2014; Tyros et al., 2016) electrical pain thresholds, (Käll et al., 2008) warm and cold detection thresholds, (Kemler et al., 2000) cold pain thresholds,

(Rushton et al., 2014) and pain drawings to evaluate widespread pain/distribution (Southerst et al., 2013).

Risk of Bias in individual and across studies

Table 3 summarises the risk of bias for individual studies categorised per outcome measure and measurement property. Overall, the risk of bias of individual studies across the modalities was rated as doubtful or inadequate. Risk of bias/overall quality across studies is summarised in table 3. Overall all outcome measures were rated as low or very low.

Results of Individual Studies

Table 2 summarises the measurement property, methodology statistical measures and results of individual studies.

TABLE 1 AND 2 HERE

Synthesis of Results

Validity

No studies were identified which evaluated content validity or criterion validity, with just one study focused on construct validity (Rushton et al., 2014).

Construct Validity

One study evaluated discriminative validity (subgroup of construct validity) (Rushton et al., 2014), assessing whether sensory evaluation (vibration perception and detection thresholds and cold pain thresholds) could discriminate between WAD and control participants. Results were that vibration detection and CPT were not supported in identification of WAD. Individual risk of bias was rated as very good and indeterminate for the COSMIN good criteria for good measurement properties. Overall, low quality evidence supports that sensory evaluation cannot discriminate between WAD and control participants.

Reliability and Measurement Error

Cold and Warm Detection Thresholds

One study evaluated intra-rater reliability of cold and warm detection thresholds using a thermosensory stimulator (Kemler et al., 2000), reporting poor reliability, more so with the

lower limb sites compared to upper limb sites. This study was rated as inadequate for risk of bias, and indeterminate on the COSMIN criteria for good measurement properties due to the statistical measures used. Very low-quality evidence overall indicates very little confidence in the reliability estimate of both cold and warm detection thresholds within a musculoskeletal trauma population. No studies evaluated measurement error for both of these outcome measures.

Electrical Pain Thresholds

One study evaluated intra-rater reliability of electrical pain thresholds using a new device called the Pain Matcher (Cefar Medical AB, Lund, Sweden) (Käll et al., 2008). Overall conclusions reported some systematic differences between scores in both sessions. Risk of bias was rated as inadequate with a rating of indeterminate on the COSMIN criteria for good measurement properties due to statistical methods used. Overall, very low-quality evidence indicates very little confidence in the reliability estimate of electrical pain thresholds using this particular device within a musculoskeletal trauma population. No studies were identified for measurement error with this outcome measure.

Pain Distribution

One study evaluated inter-rater and inter-method (paper vs electronic) reliability of pain distribution using pain drawings with overall conclusions reported as good to excellent reliability for both inter-rater and inter-method reliability (Southerst et al., 2013). Risk of bias was rated doubtful, inter-rater rated as sufficient and inter-method rated as insufficient on the COSMIN criteria for good measurement properties. Very low overall quality indicates very little confidence in the reliability estimate of pain drawings either inter-rater or inter-method within the musculoskeletal trauma population.

Measurement error was evaluated in the same study with results summarised in table 2 (Southerst et al., 2013). Risk of bias was rated as doubtful, with the COSMIN criteria for good measurement properties rated as indeterminate. The very-low overall quality indicates little confidence in the measurement error of pain drawings within the musculoskeletal trauma population.

Pinwheel

Two studies evaluated inter-rater reliability for the pinwheel (Wartenberg Pinwheel) (Bock et al., 2005), the other study did not state type of pinwheel (Bertilson et al., 2003). One reported evaluating specifically allodynia (Bock et al., 2005) with the other study reported evaluating sensitivity to pain (Bertilson et al., 2003). Although different wording to report methods was used within studies, methodology for both studies was deemed similar whereby a 'response' from the participant was sought, therefore results from the studies were narratively synthesised. Both studies reported good (Bock et al., 2005) and adequate reliability (Bertilson et al., 2003). The overall agreement was reported in one study (Bertilson et al., 2003) with >80% agreement between raters. Risk of bias was rated as doubtful for both studies (Bertilson et al., 2003; Bock et al., 2005). One study was rated sufficient (Bock et al., 2005) and one insufficient (Bertilson et al., 2003) on the COSMIN criteria for good measurement properties. Low overall quality for the pinwheel indicates limited confidence in the reliability of the pinwheel within the musculoskeletal trauma population.

The same studies investigated measurement error (Bertilson et al., 2003; Bock et al., 2005). For risk of bias, one study was rated doubtful (Bertilson et al., 2003), and one inadequate (Bock et al., 2005). Both studies were rated indeterminate for the COSMIN criteria for good measurement properties. Low overall quality indicates limited confidence in measurement error estimate for the pinwheel within the musculoskeletal trauma population was rated as low.

Pressure Pain Threshold

Two studies evaluated intra and inter-rater reliability of PPT, both reporting adequate reliability. Both used a handheld algometer (Somedic type II, Sweden) for testing. Risk of bias was rated inadequate (Prushansky et al., 2007) and doubtful (Saebo et al., 2019) for both intra and inter-rater reliability. For intra-rater reliability, with one study was rated sufficient (Prushansky et al., 2007) and one insufficient (Saebo et al., 2019) on the COSMIN criteria for good measurement properties with both studies rating sufficient for inter-rater reliability (Prushansky et al., 2007; Saebo et al., 2019). Very low overall quality for both intra and inter-rater reliability for PPT indicates very limited confidence in the reliability estimate within the musculoskeletal trauma population

The same studies investigated measurement error (Prushansky et al., 2007; Saebo et al., 2019), with results summarised in table 2. For risk of bias one study was rated inadequate (Prushansky et al., 2007) and the other study was rated doubtful (Saebo et al., 2019). Both

studies were rated as indeterminate for the COSMIN criteria for good measurement properties. Very low overall quality indicates very little confidence in measurement error estimates within the musculoskeletal trauma population.

Vibration detection Thresholds

One study evaluated intra and inter-rater reliability of vibration detection thresholds using a 128Hz tuning fork (Ragg Gardiner Brown Co) (Tyros et al., 2016). Overall conclusions were reported as excellent intra and inter-rater reliability. Risk of bias was rated as doubtful and a rating of sufficient for COSMIN criteria for good measurement properties. Very low overall quality indicates very little confidence in the reliability estimate for vibration detection thresholds within the musculoskeletal trauma population.

Measurement error was calculated in the same study including limits of agreement using Bland Altman plots, and standard error of measurement (Tyros et al., 2016). Risk of bias was rated as doubtful, and COSMIN criteria for good measurement properties rated as indeterminate. Very low overall quality indicates very little confidence in the measurement error of vibration detection thresholds within the musculoskeletal trauma population.

Responsiveness

No studies were identified which evaluated responsiveness.

TABLE 3 HERE

DISCUSSION AND CONCLUSIONS

This systematic review is the first to synthesise and evaluate outcome measures of CS and their measurement properties applied in a musculoskeletal trauma population. Stage one of this review identified 30 measures were used to evaluate CS within musculoskeletal trauma, with the majority of the research identified in populations involving WAD. The high number of outcome measures used in studies to evaluate CS in this population highlights the lack of gold standard and consensus in the literature on the most optimal outcome measure(s) for CS. Furthermore, it highlights the complexity of the concept of CS in that no measure can fully assess CS but assess clinical features such as widespread pain or secondary hyperalgesia, which are suggestive of the presence of CS only (Neblett 2018). This could

explain why multiple outcome measures which have been suggested to assess clinical features of CS e.g. QST (Cruz-Almeida and Fillingim 2014) were used in this population. A consensus of the most 'optimal' outcome measure(s) to use to detect CS would be useful in future research to allow a more standardised approach in assessing clinical features of CS. However, the current understanding of CS acknowledges that there are multiple mechanisms both peripherally and centrally which can trigger and then maintain CS (Nijs et al., 2019). Therefore, one measure or even multiple measures assessing features of CS may not be adequate to give a detailed understanding of the patient's presentation, and thus are a limitation when assessing CS.

Stage two of this systematic review evaluated measurement properties of the measures identified from stage one. From the 30 measures, measurement properties were evaluated in nine measures specifically in musculoskeletal trauma. Reliability was evaluated in seven measures – cold, warm and vibration detection thresholds, electrical and pressure pain thresholds, pain distribution and the pinwheel. Measurement error was evaluated in PPT, pain distribution and the pinwheel. Validity was evaluated in vibration disappearance and detection thresholds and CPT. No study evaluated responsiveness. PPT, the most common method found to evaluate CS, only had two studies evaluating the reliability specifically in patients following musculoskeletal trauma (Prushansky et al., 2007; Saebo et al., 2019), and CPT, the second most common method was only used within one validity study (Rushton et al., 2014). Other common outcome measures identified, such as heat pain thresholds, brachial plexus provocation test, temporal summation and central pain modulation, no studies were found evaluating measurement properties in people with musculoskeletal trauma.

Risk of bias for all studies were rated as doubtful or inadequate apart from the one study which assessed construct validity which was rated as very good. For reliability and measurement error, overall quality was rated as very low for all outcome measures apart from the pinwheel which was rated as low. For validity, despite individual methodological quality being rated as very good, overall quality was low being downgraded due to imprecision because of low sample size numbers. Low sample size, inconsistency between study results and low methodological quality was a contributing factor to low quality for reliability and measurement error.

From the results of this review, it is hard to make conclusions of the most appropriate CS outcome measure due to the high risk of bias and inconsistency of results of individual

studies. However, a body of research exists with established measurement properties particularly reliability within healthy volunteers or other musculoskeletal conditions for multiple CS outcome measures. Examples include PPT for healthy participants (Chesterton et al., 2007), healthy and acute neck pain participants (Walton et al., 2011a), and knee osteoarthritis (Wylde et al., 2011), central sensitisation index for chronic musculoskeletal conditions (Scerbo et al., 2018), and temporal summation for healthy participants (Graven-Nielsen et al., 2015; Kong et al., 2013). This has often been used as justification to use and translate measures into other populations. However, with the exception of the central sensitisation index, the risk of bias and overall quality of these studies has not been evaluated within a review. Of the reviews which do exist for reliability of thermal QST (Moloney et al., 2012) and central pain modulation (Kennedy et al., 2016), the methodological quality was variable with both reviews highlighting issues of poor reporting of blinding of raters and randomisation, and variable statistical analyses used.

This review echoes issues highlighted in previous systematic reviews in that all studies included in this review were found to have doubtful or inadequate risk of bias, with overall quality for each measure being low or very low. For individual risk of bias, reoccurring themes in all reliability studies was firstly around the reporting of methods e.g. explicit reporting of participants being stable between sessions, and the environment being similar for each session. Secondly, the time between testing sessions was variable ranging from less than a few minutes to a month. COSMIN recommends for PROMS a gap of two weeks is adequate to prevent recall bias, (Mokkink et al., 2018b) with Sim and Wright (2000) suggesting time between measurements should be large enough to ensure the measures are independent and not influenced by recall of the participant or rater, or the previous measurement. Therefore, with the nature of the measures being evaluated in this review i.e. sensory changes, a time lapse of less than five minutes could be argued to be insufficient to allow any washout period and recall from the participant. Furthermore, the variations observed in protocols of assessing clinical features of CS could potentially affect the validity and reliability and therefore the ability in drawing overall conclusions of the most appropriate measure in assessing clinical features of CS. Variations in protocols were also observed in studies in stage 1 of the review, highlighting again the lack of consistency in both measures and testing protocols when evaluating features of CS, which again could affect validity and reliability in assessing clinical features of CS.

A challenge in this review was around wording and statistical measures used to evaluate reliability and measurement error, in keeping with previous systematic reviews (Kennedy et al., 2016; Moloney et al., 2012). COSMIN recommend intra-class correlation coefficients for continuous scores and kappa coefficient statistics for dichotomous/nominal and ordinal scores when assessing reliability, (Mokkink et al., 2018b; Sim and Wright 2000) yet in some studies this was not conducted. Therefore, although some studies reported 'good reliability', overall conclusions about specific measures were difficult since inappropriate statistical analyses were used in the original studies. Additionally, the wording around reliability was challenging with multiple terms such as reproducibility, repeatability and relationship used interchangeably with reliability. This review used the COSMIN definitions and terms of measurement properties in an attempt to standardise the reporting, but it is clear further standardisation is needed in reporting reliability more generally.

Strengths and Limitations

This is the first review to evaluate CS outcome measures and their measurement properties within musculoskeletal trauma. This review followed a published protocol which was registered on PROSPERO, with a pre-defined comprehensive search strategy. Adopting a two-stage search approach allowed a comprehensive search to be conducted to identify all measures used within this population of interest. All stages of this review were conducted by two reviewers independently to limit any potential bias. Furthermore, steps were taken to ensure all authors were contacted where the population did not clearly state if the cohort had more than 90% musculoskeletal trauma.

Some limitations are recognised. In both stages, studies were excluded since they were not published in English (three articles in stage one and five articles in stage two). Therefore, it is possible that some articles evaluating CS measures were not included. Furthermore, despite efforts to clarify details to inform eligibility, a number of authors did not respond or could confirm the cohort included more than 90% musculoskeletal trauma. However, it is doubtful the overall conclusions of this review would have changed if further articles had been included. Due to the heterogeneity of the data, a meta-analysis was not possible thus a narrative review was conducted to summarise the findings. This coupled with the low quality of the included studies, made discussion and conclusions challenging.

Finally, a high number of studies within stage 1 of the review were of the WAD population. A pre-defined comprehensive search strategy was developed and piloted to

ensure all relevant articles across musculoskeletal trauma were included in addition to hand searching of key journals and grey literature. We are confident that all relevant articles have been included in this review, and this review highlights the small number of studies in other types of trauma.

Future implications

This review has highlighted two main issues:

1. There is a range of outcome measures being used to assess CS in the musculoskeletal trauma population, highlighting the complexity around CS and the evolving nature and understanding of this concept. Whilst this review does not suggest there is or should be one superior measure to use for evaluation of CS, further work towards a consensus on the most appropriate measures to assess CS within musculoskeletal trauma is needed.
2. There are a lack of studies evaluating measurement properties of CS measures within musculoskeletal trauma, and of those which have been identified, good individual methodological quality is limited with overall low quality. Further studies are warranted for multiple CS outcome measures taking into consideration consistency of terms e.g. reliability vs reproducibility, high quality methodology including appropriate statistical measures and adequate time between measurements.

Conclusion

This systematic review has identified a range of outcome measures used in musculoskeletal trauma to evaluate CS, with the majority of CS research conducted within the WAD population. Nine measures were identified with evaluated measurement properties within a musculoskeletal trauma population. Risk of bias for all studies apart from one evaluating construct validity was doubtful or inadequate, and overall quality was rated low or very low for all outcome measures and measurement properties. Conclusions about the reliability of various measures was difficult due to the wide variety of results, methods and sites tested. Further research is required to establish measurement properties within this population as well as to achieve consensus on the most appropriate measures to evaluate CS.

ETHICS AND DISSEMINATION

No research ethics was required for this systematic review due to no patient data being collected.

REFERENCES

Alqarni AM, Manlapaz D, Baxter D, Tumilty S, Mani R. Test Procedures to Assess Somatosensory Abnormalities in Individuals with Peripheral Joint Pain: A Systematic Review of Psychometric Properties. Pain Practice 2018;18: 895-924.

Bertilson BC, Grunnesjö M, Strender L. Reliability of clinical tests in the assessment of patients with neck/shoulder problems -- impact of history. Spine (03622436) 2003;28: 2222-2231.

Biurrun Manresa JA, Neziri AY, Curatolo M, Arendt-Nielsen L, Andersen OK. Test-retest reliability of the nociceptive withdrawal reflex and electrical pain thresholds after single and repeated stimulation in patients with chronic low back pain. European Journal of Applied Physiology 2011;111: 83-92.

Blasco T and Bayes R. Unreliability of the Cold Pressor Test Method in pain studies. Methods Find Exp Clin Pharmacol 1988;10: 767-772.

Bock SL, Centeno CJ, Elliott JM. The presence and interrater reliability of thoracic allodynia in a whiplash cohort. Pain Physician 2005;8: 267-270.

Carroll LJ, Holm LW, Hogg-Johnson S, Côté P, Cassidy JD, Haldeman S, Nordin M, Hurwitz EL, Carragee EJ, van der Velde G, Peloso PM, Guzman J. Course and Prognostic Factors for Neck Pain in Whiplash-Associated Disorders (WAD): Results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. European Spine Journal 2008;17: 83-92.

Chesterton LS, Sim J, Wright CC, Foster NE. Interrater reliability of algometry in measuring pressure pain thresholds in healthy humans, using multiple raters. The Clinical journal of pain 2007;23: 760-766.

Christiansen M, Zhang N, Ross M, Goodman B, Snyder CH, Smith B. How reliable are bedside measurements of sensation? Annals of Neurology 2017;82 (Supplement 21): S203.

Clark J, Nijs J, Yeowell G, Goodwin P. What Are the Predictors of Altered Central Pain Modulation in Chronic Musculoskeletal Pain Populations? A Systematic Review. Pain physician 2017;20: 487.

Clay FJ, Newstead SV, McClure RJ. A systematic review of early prognostic factors for return to work following acute orthopaedic trauma. Injury 2010;41: 787-803.

Clay FJ, Watson WL, Newstead SV, McClure RJ. A systematic review of early prognostic factors for persistent pain following acute orthopedic trauma. Pain Research & Management : The Journal of the Canadian Pain Society 2012;17: 35-44.

Coppieters I, De Pauw R, Kregel J, Malfliet A, Goubert D, Lenoir D, Cagnie B, Meeus M. Differences Between Women With Traumatic and Idiopathic Chronic Neck Pain and Women Without Neck Pain: Interrelationships Among Disability, Cognitive Deficits, and Central Sensitization. Physical Therapy 2017;97: 338-353.

Cruz-Almeida Y and Fillingim RB. Can Quantitative Sensory Testing Move Us Closer to Mechanism-Based Pain Management? Pain Medicine 2014;15: 61-72.

Cummings GS and Routan JL. Accuracy of the unassisted pain drawings by patients with chronic pain. Journal of Orthopaedic and Sports Physical Therapy 1987;8: 391-396.

Cummings GS and Routon JL. Validity of unassisted pain drawings by patients with chronic pain Physical Therapy 1985;65: 668-669.

Doménech-García V, Skuli Palsson T, Boudreau SA, Herrero P, Graven-Nielsen T. Pressure-induced referred pain areas are more expansive in individuals with a recovered fracture. PAIN 2018;159: 1972-1979.

Fingleton C, Smart K, Moloney N, Fullen BM, Doody C. Pain sensitization in people with knee osteoarthritis: a systematic review and meta-analysis. Osteoarthritis and Cartilage 2015;23: 1043-1056.

Fricton J and Schiffman E. A pressure algometer for muscle palpation - reliability and validity. Journal of Dental Research 1986;65: 334-334.

Graven-Nielsen T and Arendt-Nielsen L. Assessment of mechanisms in localized and widespread musculoskeletal pain. Nature Reviews Rheumatology 2010;6: 599-606.

Graven-Nielsen T, Vaegter HB, Finocchietti S, Handberg G, Arendt-Nielsen L. Assessment of musculoskeletal pain sensitivity and temporal summation by cuff pressure algometry: a reliability study. Pain 2015;156: 2193-2202.

Harte SE, Harris RE, Clauw DJ. The neurobiology of central sensitization. Journal of Applied Biobehavioral Research 2018;23: e12137.

IASP. International Association for the Study of Pain: IASP Terminology. 2017; Available from: https://www.iasp-pain.org/terminology?navItemNumber=576.

Käll LB, Kowalski J, Stener-Victorin E. Assessing pain perception using the PainMatcher® in patients with whiplash-associated disorders. Journal of rehabilitation medicine 2008;40: 171-177.

Katz J and Seltzer Ze. Transition from acute to chronic postsurgical pain: risk factors and protective factors. Expert Review of Neurotherapeutics 2009;9: 723-744.

Kemler MA, Reulen JP, van Kleef M, Barendse GA, van den Wildenberg FA, Spaans F. Thermal thresholds in complex regional pain syndrome type I: sensitivity and repeatability of the methods of limits and levels. Clinical Neurophysiology 2000;111: 1561-1568.

Kennedy DL, Kemp HI, Ridout D, Yarnitsky D, Rice ASC. Reliability of conditioned pain modulation: a systematic review. PAIN 2016;157: 2410-2419.

Kong J-T, Johnson KA, Balise RR, Mackey S. Test-retest reliability of thermal temporal summation using an individualized protocol. The Journal of Pain 2013;14: 79-88.

Lahoda R, Stacher G, Bauer P. Experimentally induced pain: measurement of pain threshold and pain tolerance using a new apparatus for electrical stimulation of the skin. International journal of clinical pharmacology and biopharmacy 1977;15: 51-56.

Latremoliere A and Woolf CJ. Central Sensitization: A Generator of Pain Hypersensitivity by Central Neural Plasticity. The Journal of Pain 2009;10: 895-926.

Margolis RB, Chibnall JT, Tait RC. Test-retest reliability of the pain drawing instrument. Pain 1988;33: 49-51.

Margolis RB, Tait RC, Krause SJ. A rating system for use with patient pain drawings. Pain 1986;24: 57-65.

Middlebrook N, Rushton AB, Heneghan NR, Falla D. Measures of central sensitisation and their measurement properties in the adult musculoskeletal trauma population: a protocol for a systematic review and data synthesis. BMJ Open 2019;9: e023204.

Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA Statement. BMJ 2009;339.

Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, Terwee CB. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Qual Life Res 2018a;27: 1171-1179.

Mokkink LB, Prinsen C, Patrick DL, Alonso J, Bouter LM, de Vet H, Terwee CB, Mokkink L. COSMIN methodology for systematic reviews of Patient- Reported outcome measures (PROMs): User Manual. 2018b; Available from: https://www.cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018.pdf.

Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HCW. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research 2010a;19: 539-549.

Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HCW. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. Journal of Clinical Epidemiology 2010b;63: 737-745.

Moloney NA, Hall TM, Doody CM. Reliability of thermal quantitative sensory testing: a systematic review. Journal of rehabilitation research and development 2012;49: 191.

Myburgh C, Lauridsen HH, Larsen AH, Hartvigsen J. Standardized manual palpation of myofascial trigger points in relation to neck/shoulder pain; the influence of clinical experience on inter-examiner reproducibility. Manual Therapy 2011;16: 136-140.

Neblett R. The central sensitization inventory: A user's manual. Journal of Applied Biobehavioral Research 2018;23: e12123.

NICE. Guideline Scope: Rehabilitation after traumatic injury. 2018; Available from: https://www.nice.org.uk/guidance/gid-ng10105/documents/final-scope.

Nijs J, Leysen L, Vanlauwe J, Logghe T, Ickmans K, Polli A, Malfliet A, Coppieters I, Huysmans E. Treatment of central sensitization in patients with chronic pain: time for change? Expert Opinion on Pharmacotherapy 2019;20: 1961-1970.

Nijs J, Torres-Cueco R, van Wilgen CP, Girbes EL, Struyf F, Roussel N, van Oosterwijck J, Daenen L, Kuppens K, Vanwerweeen L, Hermans L, Beckwee D, Voogt L, Clark J, Moloney N, Meeus M. Applying modern pain neuroscience in clinical practice: criteria for the classification of central sensitization pain. Pain Physician 2014;17: 447-457.

O'Neill S, Graven-Nielsen T, Manniche C, Arendt-Nielsen L. Reliability and validity of a simple and clinically applicable pain stimulus: Sustained mechanical pressure with a spring-clamp. Chiropractic and Manual Therapies 2014;22 (1) (no pagination).

Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, Terwee CB. COSMIN guideline for systematic reviews of patient-reported outcome measures. Quality of Life Research 2018.

Prushansky T, Handelzalts S, Pevzner E. Reproducibility of pressure pain threshold and visual analog scale findings in chronic whiplash patients. Clinical Journal of Pain 2007;23: 339-345.

Reigo T, Tropp H, Timpka T. Pain drawing evaluation--the problem with the clinically biased surgeon. Intra- and interobserver agreement in 50 cases related to clinical bias. Acta Orthopaedica Scandinavica 1998;69: 408-411.

Ris I, Barbero M, Falla D, Larsen M, Kraft MN, Søgaard K, Juul-Kristensen B. Pain extent is more strongly associated with disability, psychological factors, and neck muscle function in people with non-traumatic versus traumatic chronic neck pain: a cross sectional study. European journal of physical and rehabilitation medicine 2018.

Rivara FP, MacKenzie EJ, Jurkovich GJ, Nathens AB, Wang J, Scharfstein DO. Prevalence of pain in patients 1 year after major trauma. Archives of Surgery 2008;143: 282-287.

Rushton A, Wright C, Kontakiotis N, Mystrakis A, Frydas D, Heneghan N. Discriminative validity of sensory evaluation in a whiplash-associated disorder II population. International Journal of Therapy & Rehabilitation 2014;21: 460-467.

Saebo H, Naterstad IF, Stausholm MB, Bjordal JM, Joensen J. Reliability of pain pressure threshold algometry in persons with conservatively managed wrist fractures. Physiotherapy research international : the journal for researchers and clinicians in physical therapy 2019: e1797-e1797.

Scerbo T, Colasurdo J, Dunn S, Unger J, Nijs J, Cook C. Measurement Properties of the Central Sensitization Inventory: A Systematic Review. Pain Practice 2018;18: 544-554.

Sim J and Wright C. Research in health care: concepts, designs and methods. Hampshire, UK: Nelson Thornes. 2000.

Southerst D, Stupar M, Côté P, Mior S, Stern P. The reliability of measuring pain distribution and location using body pain diagrams in patients with acute whiplash-associated disorders. Journal of manipulative and physiological therapeutics 2013;36: 395-402.

Starz TW, Sinclair JD, Okifuji A, Turk DC, McConnell R. Interrater reliability of a standardized manual tender point examination protocol Arthritis and Rheumatism 1995;38: 994-994.

Sterling M, Jull G, Vicenzino B, Kenardy J. Sensory hypersensitivity occurs soon after whiplash injury and is associated with poor recovery. Pain 2003;104: 509-517.

Tyros I, Soundy A, Heneghan NR. Vibration sensibility of the median nerve in a population with chronic whiplash associated disorder: Intra- and inter-rater reliability study. Manual Therapy 2016;25: 81-86.

Van Oosterwijck J, Nijs J, Meeus M, Paul L. Evidence for central sensitization in chronic whiplash: A systematic literature review. European Journal of Pain 2013;17: 299-312.

Vardeh D, Mannion RJ, Woolf CJ. Toward a Mechanism-Based Approach to Pain Diagnosis. The Journal of Pain 2016;17: T50-T69.

Vuilleumier PH, Biurrun Manresa JA, Ghamri Y, Mlekusch S, Siegenthaler A, Arendt-Nielsen L, Curatolo M. Reliability of Quantitative Sensory Tests in a Low Back Pain Population. Regional Anesthesia & Pain Medicine 2015;40: 665-673.

Walton D, MacDermid J, Nielson W, Teasell R, Chiasson M, Brown L. Reliability, standard error, and minimum detectable change of clinical pressure pain threshold testing in people with and without acute neck pain. Journal of Orthopaedic & Sports Physical Therapy 2011a;41: 644-650.

Walton D, MacDermid J, Nielson W, Teasell R, Reese H, Levesque L. Pressure pain threshold testing demonstrates predictive ability in people with acute whiplash. Journal of Orthopaedic & Sports Physical Therapy 2011b;41: 658-665.

Williams DA. Phenotypic features of central sensitization. Journal of Applied Biobehavioral Research 2018;23: e12135.

Williamson OD, Epi GDC, Gabbe BJ, Physio B, Cameron PA, Edwards ER, Richardson MD, Group obotVOTORP. Predictors of Moderate or Severe Pain 6 Months After Orthopaedic Injury: A Prospective Cohort Study. Journal of Orthopaedic Trauma 2009;23: 139-144.

Woolf CJ. Central sensitization: Implications for the diagnosis and treatment of pain. PAIN 2011;152: S2-S15.

Woolf CJ. Pain amplification—A perspective on the how, why, when, and where of central sensitization. Journal of Applied Biobehavioral Research 2018;23: e12124.

Wylde V, Palmer S, Learmonth ID, Dieppe P. Test–retest reliability of Quantitative Sensory Testing in knee osteoarthritis and healthy participants. Osteoarthritis and Cartilage 2011;19: 655-658.

Author's Contributions

NM is PhD student and DF (lead supervisor), AR and NH are supervisors. NM drafted the initial version of the manuscript with DF, AR and NH all providing guidance on topic, methodology and analyses. PK was the second reviewer for stage 1 and DA the second reviewer for stage 2. All authors reviewed and commented on each draft of the protocol. All authors have approved the final manuscript.

Table 1. Study characteristics of included studies in stage two

Study	Country	Study Design	Sample Size	Participant Characteristics	Trauma	MOI	Duration	CS Outcome Measure
Bertilson et al., (2003)	Sweden	Clinical Trial	n=100	Age: Mean (range) Without history: 42.7 (18-66) With history: 43.5 (25-66) Gender: M/F Group 1: 20/30 Group 2: 13/37	Trauma over 90% confirmed by author	Not reported	5 days-60 years	Sensitivity to Pain - pinwheel
Bock et al., (2005)	USA	Prospective case series	n=22	Age: Mean (SD), range 40.9 (14.8), 16-72 Gender: M/F 5/17	Whiplash	Motor vehicle accident	0.25-1.25 years	Allodynia - Wartenberg pinwheel
Käll et al.,	Sweden	Secondary	n=47	Age: Mean (range)	Whiplash	Not reported	42-121 days	Electrical pain

(2008)		analysis of RCT		31 (18-61) Gender: M/F 17/30				threshold
Kemler et al., (2000)	Netherlands	Not reported	n=53 Foot group n=20 Hand group n=33	Age: Mean (range) 38.6 (21-65) Gender: M/F 16/37	Complex regional pain syndrome type I – confirmed 90% caused by trauma by author	Not reported	9-120 months	Warm detection threshold, Cold detection threshold
Prushansky et al., (2007)	Israel	Not reported	n=21	Age: Mean (SD), range 40.5 (12.3), 18-64 Gender: M/F 8/13	Whiplash II	Road traffic collision	6-132 months	Pressure pain thresholds
Rushton et al., (2014)	United Kingdom	Case control, observational	n=42 Whiplash	Age: Median (IRQ), range Whiplash 28.5 (12.8), 20-	Whiplash II	Not reported	2-66 months	Vibration perception threshold,

			n=20 Control n=22	55 Control 26 (4), 20-39 Gender: M/F Whiplash 7/13 Control 11/9				vibration disappearance threshold, cold pain threshold
Saebo et al., (2019)	Norway	Cross sectional	n=75	Age Mean (SD), range 56 (21), 18-97) Gender: M/F 19/56	Wrist fractures	Not reported	Post cast removal	Pressure pain thresholds
Southerst et al., (2013)	Canada	Reliability study nested within RCT	n=80	Age: Mean 39.1 Gender M/F 24/56	Whiplash I or II	Road traffic collision	1-375 days	Pain Distribution

Tyros et al., (2016)	United Kingdom	Double blinded cross-sectional study	n=26	Age: Mean (SD) 29.9 (10) Gender: M/F 8/18	Whiplash II	Not reported	>6 months	Vibration disappearance threshold

F, Female; IRQ, Interquartile Range; M, Male; MOD, mechanism of injury SD, Standard Deviation; RCT, Randomised Controlled Trial

References

Bertilson BC, Grunnesjö M, Strender L. Reliability of clinical tests in the assessment of patients with neck/shoulder problems -- impact of history. Spine (03622436) 2003;28: 2222-2231.

Bock SL, Centeno CJ, Elliott JM. The presence and interrater reliability of thoracic allodynia in a whiplash cohort. Pain Physician 2005;8: 267-270.

Käll LB, Kowalski J, Stener-Victorin E. Assessing pain perception using the PainMatcher® in patients with whiplash-associated disorders. Journal of rehabilitation medicine 2008;40: 171-177.

Kemler MA, Reulen JP, van Kleef M, Barendse GA, van den Wildenberg FA, Spaans F. Thermal thresholds in complex regional pain syndrome type I: sensitivity and repeatability of the methods of limits and levels. Clinical Neurophysiology 2000;111: 1561-1568.

Prushansky T, Handelzalts S, Pevzner E. Reproducibility of pressure pain threshold and visual analog scale findings in chronic whiplash patients. Clinical Journal of Pain 2007;23: 339-345.

Rushton A, Wright C, Kontakiotis N, Mystrakis A, Frydas D, Heneghan N. Discriminative validity of sensory evaluation in a whiplash-associated disorder II population. International Journal of Therapy & Rehabilitation 2014;21: 460-467.

Saebo H, Naterstad IF, Stausholm MB, Bjordal JM, Joensen J. Reliability of pain pressure threshold algometry in persons with conservatively managed wrist fractures. Physiotherapy research international : the journal for researchers and clinicians in physical therapy 2019: e1797-e1797.

Southerst D, Stupar M, Côté P, Mior S, Stern P. The reliability of measuring pain distribution and location using body pain diagrams in patients with acute whiplash-associated disorders. Journal of manipulative and physiological therapeutics 2013;36: 395-402.

Tyros I, Soundy A, Heneghan NR. Vibration sensibility of the median nerve in a population with chronic whiplash associated disorder: Intra- and inter-rater reliability study. Manual Therapy 2016;25: 81-86.

Table 2. Summary of measurement property results of included studies in stage two

Study	Measurement Property	Raters & Testing schedule	Groups	Sites Tested	Statistical measures	Results
Bertilson et al., (2003)	Reliability (inter-rater)	2 raters Same day testing Time Interval: not reported	With clinical history Without clinical history	Within dermatome of: Chin (C2) Neck (C3) Shoulder (C4) Upper arm (C5) Thumb (C6) Middle finger (C7) Little finger (C8) Axilla (T2) Chest (T4) Foot (L5)	Overall agreement % Kappa coefficients	Overall agreement: Without history: 81% With history: 84% Kappa coefficient (SD): Without history 0.57 (0.12) With history 0.67 (0.11)
Bock et al.,	Reliability (inter-	2 raters	1 group	T1-T12 spinal	Kappa coefficients	Kappa (95% CI)

(2005)	rater)	Same day testing Time interval: within 5 minutes		processes		0.8039 (0.7465,0.8163)
Käll et al., (2008)	Reliability (intra-rater)	1 rater Same-day testing Time Interval: 5 minutes	1 group	Device placed between right thumb and index fingers	Relative rank variance (RV) Relative concentration (RC) Relative position (RP)	RV (SE, 95% CI): 0.10 (0.06, 0.00,0.22) RC (SE, 95% CI) 0.07 (0.06, 0.18, 0.05) RP (SE, 95% CI) 0.16 (0.05, 0.25, -0.07)
Kemler et al., (2000)	Reliability (Intra-rater)	1 rater Between day testing	1 group consisting of CRPS with either foot or hand.	Foot and wrist bilaterally depending on CRPS presentation	Coefficient of repeatability	Cold perception threshold MLE Unaffected wrist 0.8 Affected wrist 0.7

		Time Interval: 1 month	2 methods assessed: MLE & MLI.			Unaffected foot 4.1 Affected foot 5.8 MLI Unaffected wrist 2.3 Affected wrist 3.7 Unaffected foot 5.3 Affected foot 3.4 Warm detection Threshold MLE Unaffected wrist 1.0 Affected wrist 2.0 Unaffected foot 5.4 Affected foot 4.2 MLI Unaffected wrist 1.7

						Affected wrist 5.0 Unaffected foot 2.9 Affected foot 4.4
Prushansky et al., (2007)	Reliability (Intra and Inter-rater)	2 raters Inter rater: same day Time interval: 15 minutes Intra-rater: Mean (SD) 7.9 (1.9) days after test 1	1 group	C2, C4, C6 bilaterally	Pearson's product moment correlation coefficient (Pearson r) ICC $_{(2,k)}$ SEM SRD LOA's	Intra-rater Pearson r C2 R=0.76 L=0.78 C4 R=0.83 L=0.85 C6 R=0.82 L=0.81 ICC C2 R=0.85 L=0.86 C4 R=0.9 L=0.91 C6 R=0.9 L=0.89 SEM/SRD (kPa) C2 R=15.3/42.4 L=14.5/40.2 C4 R=17.6/48.7 L=16.7/46.3 C6 R=21.3/46.3 L=21.7/58.4

						Inter-rater (ICC) C2 R=0.88 L=0.97 C4 R=0.90 L=0.93 C6 R=0.97 L=0.96
Rushton et al., (2014)	Discriminative (construct) validity	1 session	Whiplash and control group	Vibration thresholds – thenar and hypothenar eminence CPT's thenar eminence, dorsal aspect of 5^{th} metacarpal and remote site of mid cervical spine	Associations- Kendall's tau Logistic regression/ Hosmer-Lemeshow test.	Associations: Moderate association of vibration threshold and local sites Moderate to very high association between CPT and thenar eminence VT or CPT at thenar eminence discriminated between groups Hosmer-Lemeshow X^2 (df,p) VT 4.311 (8, .828) CPT 3.432 (8, .904)
Saebo et al., (2019)	Reliability (Intra and Inter-rater)	3 raters (A, B, C)	1 group	Dorsal side of radius, mid position perpendicular to healing fracture	ICC $_{(1,1)}$ (95% CI)	Intra-rater Injured Wrist Rater A $ICC_{(1,1)}$=0.825 (0.737,0.886) ICC $_{(3,1)}$

		Same day Time Interval: 3-5 minutes		line. Mimicked position on injured side	ICC $_{(3,1)}$ (95% CI) MDC	=0.824 (0.735,0.885). MDC=63.3 Rater B $ICC_{(1,1)}$= 0.640 (0.444,0.778) $ICC_{(3,1)}$=0.636 (0.437,0.776). MDC=138.5 Rater C $ICC_{(1,1)}$= 0.860 (0.711, 0.935) $ICC_{(3,1)}$=0.855 (0.698,0.933). MDC=78.9 Non-Injured Wrist Rater A $ICC_{(1,1)}$=0.765 (0.653,0.845) $ICC_{(3,1)}$=0.776 (0.640,0.868). MDC=98.8 Rater B $ICC_{(1,1)}$=0.667 (0.480,0.796) $ICC_{(3,1)}$=0.669 (0.482,0.798). MDC=162.8 Rater C $ICC_{(1,1)}$=0.843 (0.679,0.927) $ICC_{(3,1)}$=0.841 (0.672,0.927). MDC=86.5 Inter Rater Injured Wrist Rater A-B $ICC_{(1,1)}$=0.617 (0.413,0.763) $ICC_{(3,1)}$=0.778 (0.640,0.868). MDC=120.1 Rater A-C $ICC_{(1,1)}$=0.706 (0.443, 0.858) $ICC_{(3,1)}$=0.737 (0.488,0.875). MDC=111.8 Non Injured Wrist Rater A-B $ICC_{(1,1)}$=0.551 (0.326, 0.717) $ICC_{(3,1)}$=0.585 (0.369,0.741). MDC=157.7 Rater A-C $ICC_{(1,1)}$=0.710 (0.450, 0.860)

						$ICC_{(3,1)}$=0.714 (0.451, 0.863). MDC=114.6
Southerst et al., (2013)	Reliability (inter-rater and inter-method)	2 raters Same day Time Interval: drawings completed ,1 minute after	1 group	N/A	$ICC_{(2,1)}$ (95% CI) LOA using Bland Altman Plots	Inter-rater Paper version ICC (CI) = 0.925 (0.901,0.946) Electronic version ICC (CI) = 0.997 (0.995,0.998) Mean difference and LOA: Paper version: -0.56 +/- 6.5% Electronic version: -1.18 +/- 7.8% Inter-method Examiner 1 ICC=0.63 (0.54,0.72) Examiner 2 ICC=0.90 (0.87,0.93) Mean difference and LOA: -Examiner 1 0.51 +/- 11.7% Examiner 2 0.19 +/- 13.3%
Tyros et al., (2016)	Reliability (intra and inter-rater)	2 raters Same day	1 group	Thenar eminence	$ICC_{(2,1)}$ (95% CI) SEM	Intra-rater: Rater 1 ICC=0.972 Rater 2 ICC=0.955

		Time Interval: 1 minute between 3 measurements for intra-rater, 5 minutes between raters			LOA using Bland Altman Plots	Inter-rater: ICC=0.983 (0.971, 0.991) SEM 0.358 True SEM 0.702 LOA Upper limit 4.415. Lower Limit -2.899

C, Cervical; CI, Confidence Interval; CPT, Cold Pain Threshold; df; degrees of freedom for wald test, ICC, Intraclass correlation coefficient; LOA, Limits of Agreement; L, lumbar; MDC, Minimal Detectable Change; MLE, method of levels; MLI, method of limits; SE, Standard Error; SEM, Standard Error of Measurement; SRD, Smallest Real Difference; T, thoracic; VT, Vibration threshold

References

Bertilson BC, Grunnesjö M, Strender L. Reliability of clinical tests in the assessment of patients with neck/shoulder problems -- impact of history. Spine (03622436) 2003;28: 2222-2231.

Bock SL, Centeno CJ, Elliott JM. The presence and interrater reliability of thoracic allodynia in a whiplash cohort. Pain Physician 2005;8: 267-270.

Käll LB, Kowalski J, Stener-Victorin E. Assessing pain perception using the PainMatcher® in patients with whiplash-associated disorders. Journal of rehabilitation medicine 2008;40: 171-177.

Kemler MA, Reulen JP, van Kleef M, Barendse GA, van den Wildenberg FA, Spaans F. Thermal thresholds in complex regional pain syndrome type I: sensitivity and repeatability of the methods of limits and levels. Clinical Neurophysiology 2000;111: 1561-1568.

Prushansky T, Handelzalts S, Pevzner E. Reproducibility of pressure pain threshold and visual analog scale findings in chronic whiplash patients. Clinical Journal of Pain 2007;23: 339-345.

Rushton A, Wright C, Kontakiotis N, Mystrakis A, Frydas D, Heneghan N. Discriminative validity of sensory evaluation in a whiplash-associated disorder II population. International Journal of Therapy & Rehabilitation 2014;21: 460-467.

Saebo H, Naterstad IF, Stausholm MB, Bjordal JM, Joensen J. Reliability of pain pressure threshold algometry in persons with conservatively managed wrist fractures. Physiotherapy research international : the journal for researchers and clinicians in physical therapy 2019: e1797-e1797.

Southerst D, Stupar M, Côté P, Mior S, Stern P. The reliability of measuring pain distribution and location using body pain diagrams in patients with acute whiplash-associated disorders. Journal of manipulative and physiological therapeutics 2013;36: 395-402.

Tyros I, Soundy A, Heneghan NR. Vibration sensibility of the median nerve in a population with chronic whiplash associated disorder: Intra- and inter-rater reliability study. Manual Therapy 2016;25: 81-86.

Table 3 Summary of risk of bias, criteria for good measurement properties and overall quality of evidence (GRADE)

Measurement property outcome measure	Study	Risk of Bias	Criteria of good measurement properties	Overall Rating	Quality of evidence
Reliability					
Cold Detection Thresholds Intra-rater	Kemler et al., (2000)	Inadequate	?	?	Very low
Electrical Pain Thresholds Intra-rater	Käll et al., (2008)	Inadequate	?	?	Very low
Pain Distribution Inter-method	Southerst et al., (2013)	Doubtful	-	-	Very Low
Pain Distribution Inter-rater	Southerst et al., (2013)	Doubtful	+	+	Very Low
Pinwheel Inter-rater	Bertilson et al., (2003)	Doubtful	-	+	Low
	Bock et al., (2005)	Doubtful	+		
Pressure Pain Thresholds Intra-rater	Prushansky et al., (2007)	Inadequate	+	+	Very low
	Saebo et al., (2019)	Doubtful	-		
Pressure Pain Thresholds Inter-rater	Prushansky et al., (2007)	Inadequate	+	+	Very low
	Saebo et al., (2019)	Doubtful	+		

Vibration Detection Thresholds Intra-rater	Tyros et al., (2016)	Doubtful	+	+	Very low
Vibration Detection Thresholds Inter-rater	Tyros et al., (2016)	Doubtful	+	+	Very low
Warm detection Thresholds Intra-rater	Kemler et al., (2000)	Inadequate	?	?	Very low
Measurement Error					
Pain Distribution	Southerst et al., (2013)	Doubtful	?	?	Very Low
Pinwheel	Bertilson et al., (2003)	Doubtful	?	?	Low
	Bock et al., (2005)	Inadequate	?		
Pressure Pain Thresholds	Prushansky et al., (2007)	Inadequate	?	?	Very Low
	Saebo et al., (2019)	Doubtful	?		
Vibration detection thresholds	Tyros et al., (2016)	Doubtful	?	?	Very Low
Validity					
Construct Validity	Rushton et al., (2014)	Very Good	?	?	Low

+, sufficient; -, insufficient; ?, indeterminate, +, inconsistent

References

Bertilson BC, Grunnesjö M, Strender L. Reliability of clinical tests in the assessment of patients with neck/shoulder problems -- impact of history. Spine (03622436) 2003;28: 2222-2231.

Bock SL, Centeno CJ, Elliott JM. The presence and interrater reliability of thoracic allodynia in a whiplash cohort. Pain Physician 2005;8: 267-270.

Käll LB, Kowalski J, Stener-Victorin E. Assessing pain perception using the PainMatcher® in patients with whiplash-associated disorders. Journal of rehabilitation medicine 2008;40: 171-177.

Kemler MA, Reulen JP, van Kleef M, Barendse GA, van den Wildenberg FA, Spaans F. Thermal thresholds in complex regional pain syndrome type I: sensitivity and repeatability of the methods of limits and levels. Clinical Neurophysiology 2000;111: 1561-1568.

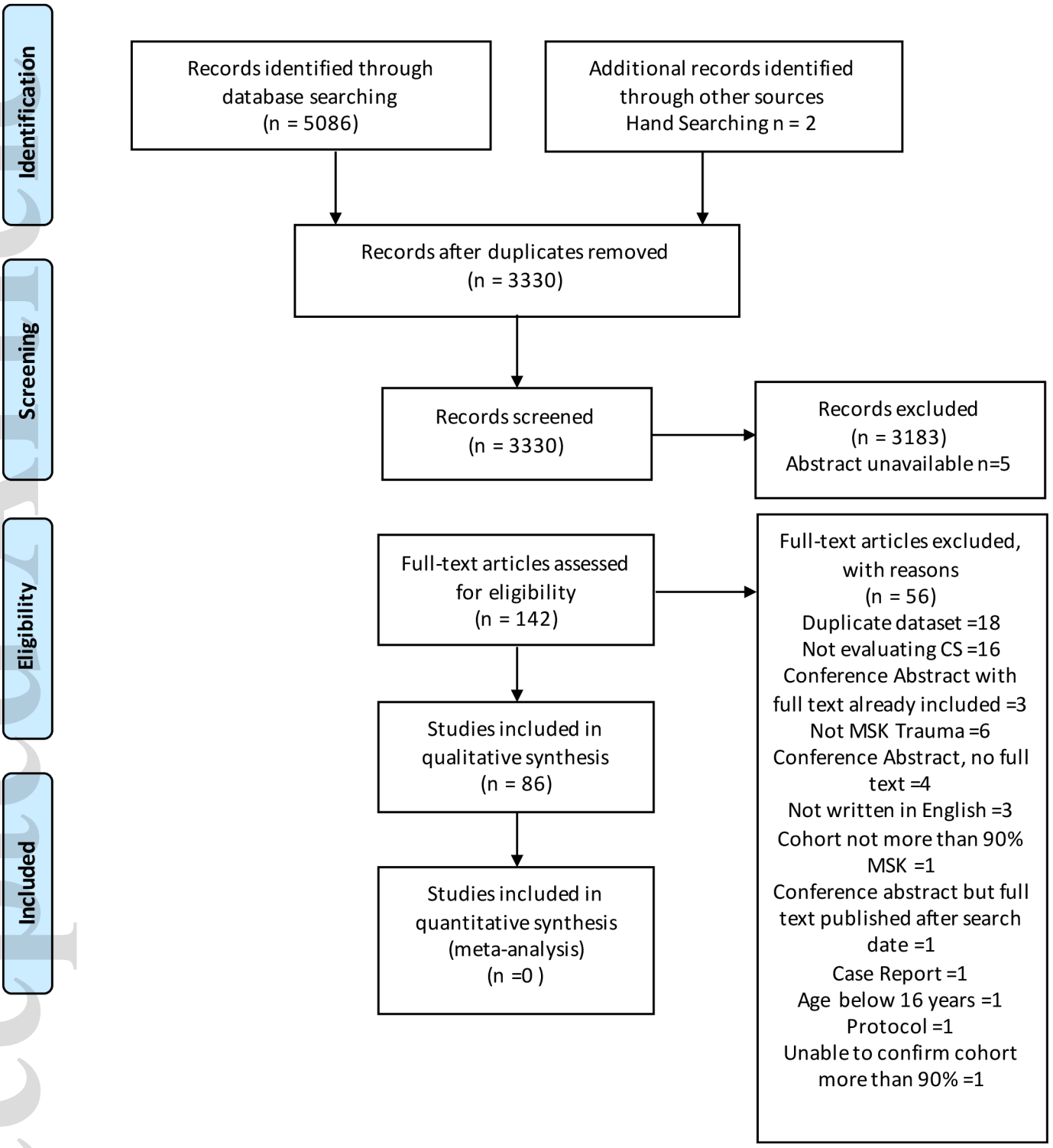
Prushansky T, Handelzalts S, Pevzner E. Reproducibility of pressure pain threshold and visual analog scale findings in chronic whiplash patients. Clinical Journal of Pain 2007;23: 339-345.

Rushton A, Wright C, Kontakiotis N, Mystrakis A, Frydas D, Heneghan N. Discriminative validity of sensory evaluation in a whiplash-associated disorder II population. International Journal of Therapy & Rehabilitation 2014;21: 460-467.

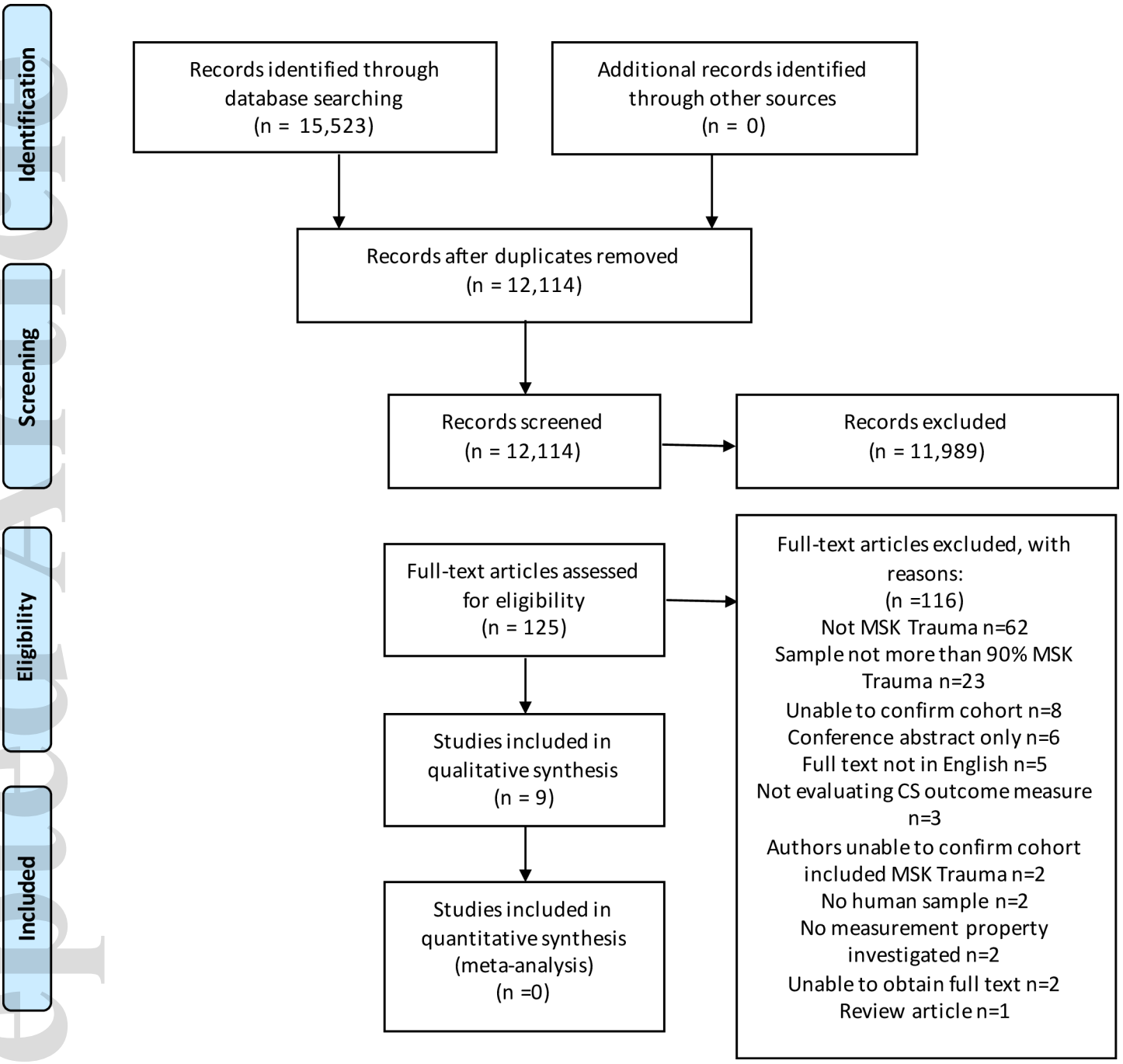
Saebo H, Naterstad IF, Stausholm MB, Bjordal JM, Joensen J. Reliability of pain pressure threshold algometry in persons with conservatively managed wrist fractures. Physiotherapy research international : the journal for researchers and clinicians in physical therapy 2019: e1797-e1797.

Southerst D, Stupar M, Côté P, Mior S, Stern P. The reliability of measuring pain distribution and location using body pain diagrams in patients with acute whiplash-associated disorders. Journal of manipulative and physiological therapeutics 2013;36: 395-402.

Tyros I, Soundy A, Heneghan NR. Vibration sensibility of the median nerve in a population with chronic whiplash associated disorder: Intra- and inter-rater reliability study. Manual Therapy 2016;25: 81-86.



ejp_1670_f1.tiff



ejp_1670_f2.tiff